NEJLT

Northern European Journal of Language Technology

www.nejlt.org

Volume 7, December 2021
ISSN 2000-1553

NEJLT Editorial Team 2021

6 Questions for Socially Aware Language Technologies

Diyi Yang, Georgia Institute of Technology, Georgia, USA dyang888@gatech.edu

Abstract Over the last few decades, natural language processing (NLP) has dramatically improved performance and produced industrial applications like personal assistants. Despite being sufficient to enable these applications, current NLP systems largely ignore the social part of language. This severely limits the functionality and growth of these applications. This letter discusses 6 questions towards how to build socially aware language technologies, with the hope of stimulating discussion, inspiring more research into Social NLP, and pushing our research field to the next level.

Over the last few decades, natural language processing (NLP) has had increasing success, and has dramatically improved performance and produced industrial applications like machine translation, search, and personal assistants. The recent Generative Pre-trained Transformer 3 (GPT3) learns language through exposure to numerous examples of data more than a human can access during their life (Brown et al., 2020) , and exhibits state of art performances on a wide range of tasks and their zero-shot learning settings.

Despite being sufficient for applications mentioned above, the current NLP systems largely ignore the social part of language, i.e., they pay attention only to what is said, but not to who says it, in what context and for what goals. This limitation severely limits both the functionality and growth of these applications, such as conversational agents' inconsistent personality and incoherent argument when conducting dialogues with humans, the failure of machine translation in generating culturally respectful outputs, the inability of current systems in commonsense reasoning, or the general struggles of current systems with social intelligence. Ultimately, the goal of NLP is to understand language the way a human does. However, it seems hard to argue that NLP models have reached human level capacity, as language is more than just information —it is about how human use a complex system of words, structures and grammar to effectively communicate with others.

I argue that getting the content correct is not enough and we should push forward on how to build socially aware language technologies that can understand and model social factors in language, interpersonal relations around language use, and the context where language is being used. The idea of language as a social construct is not new: linguistics and philosophy have long modeled it this way (Wittgenstein, 2010; Eckert, 2012) . For instance, instead of pure syntax and semantics, systemic functional linguistics (SFL) (Halliday and Matthiessen, 2013) studies language and its functions in social settings. Grice (1975) laid out four maxims that govern effective communication in social situations, *quality*—make your contribution true, do not lie or make unsupported claims, *quantity*—make your contribution as informative as is required (but not more informative), *relevance*, and *manner*—be brief and orderly and avoid obscurity of expression and ambiguity. Our recent work introduced a set of seven social factors in language and their use in NLP (Hovy and Yang, 2021); Nguyen et al. (2021) highlighted ways of learning and representing social meaning in NLP. These frameworks are quite useful in highlighting and formalizing socially aware language technologies, though there are still obstacles. Overall, I envision 6 questions about socially aware language technologies that need to be thought clearly in order to push our field to the next level.

(1) Is theory necessary in the age of data? Deeper and larger neural networks learn over massive amount of data in an end-to-end way. Should social NLP be model and data oriented? Subtle social factors are often difficult to be defined and measured, especially when "*what is said is not what is meant*", such as sarcasm, irony, deception, and any other situation that requires a "social" interpretation. For that, we need *good* theories to characterize these language phenomena, such as using the aforementioned SFL and social factors taxonomy, as well as social or linguistic theories related to individual NLP phenomena like Brown and Levinson's politeness theory (Brown and Levinson, 1987) and the incongruity theory behind humor (Lefcourt, 2001). Such theories provide grounded perspectives of linking social and language phenomena, towards the knowledge we are advancing. On the other hand, data can

extend and inform theory, as many prior theories were produced by speakers of a small set of European languages in a narrow social class stratum, with a dearth of exposure to a variety of utterances.

(2) Are benchmarks the right way to go? One key assumption of most NLP tasks is to reason over the provided benchmark. However, a single corpus might not be enough to include and represent the dynamic scenarios associated with a social phenomenon in terms of its size, genre, and population. For instance, compositionality, commonsense, or implications are key to our daily interactions, but it is often difficult to collect these rich natural situations. "*The abilities of a four-year-old that we take for granted ... answering a question*" (Pinker, 2003) do not require enormous computational or data resources to be achieved. How can we enable models to conduct open domain understanding, as social intelligence goes beyond a fixed corpus? Insights from efforts such as Kiela et al. (2021) and Bowman and Dahl (2021) and perhaps **living** benchmarks via crowdsourcing that can grow and allow for flexible input or output could help benchmark social NLP.

(3) Should social NLP models passively learn or proactively experience the world? Current NLP systems that take social factors in account mainly have been using observational data from online media or other user generated data, though there are a few exceptions in actively simulating data. This "passive" fashion only allows models to examine what is in the data and learn from it from an *association* perspective, but not easily adaptable to new scenarios even with simpler tasks. Moravec's paradox (Moravec, 1988) stated that *"it is comparatively easy to make computers exhibit adult level performance on intelligence tests..., and difficult or impossible to give them the skills of a one-year-old when it comes to perception"*. Perhaps one direction is to let NLP systems experience the world and learn, adapt their use from interacting with human. In practice, experience or interaction involves more than exchanging information via language, but also a wide range of aspects related to social and interpersonal factors reflected in rich modalities. When proactively experiencing the world, socially aware NLP also needs to go beyond text to adequately model the complexity for better understandings of language use.

(4) Should the model and evaluation stay the same? Subtle social factors are often hard to be scaled for annotation due to its subjective nature, and social scenarios often produce dynamically changing data. These *socially low resourced and evolving scenarios* poses new challenges for modern neural network techniques, making it hard for gigantic models to comprehend the world reasonably. For instance, GPT3's less encouraging results when it comes to talking about COVID-19 in late 2020 or historical figures such as asking Steve Jobs GPT-3 "*where are you right now*" and being replied as "*I'm inside Apple's headquarters in Cupertino, California*"—coherent but hardly an up-to-date/trustworthy one (Vincent, 2021). This calls for advances in methodologies that can learn with limited data and evolving facts. Not only with models, especially when social factors are involved, it might be intractable to evaluate such systems (Paullada et al., 2020; Flek, 2020). Current NLP models often use deterministic assessments to compare to some standards or ground-truths. However, these may be inadequate in capturing the nuances of social NLP, as there may be little to none ground-truth, and outputs can be various and change depending on the speaker, receiver, or other aforementioned social factors. Discrepancies with ground-truth might still be acceptable, but could also be detrimental when it comes to high stakeholder scenarios, such as inappropriate outputs from chatbots in counselling context.

(5) How can social NLP be responsible and reproducible? Unique bottlenecks for responsible social NLP includes data collection, and the associated questions about privacy, protection, and ethics, all of which we need to be aware of for doing the right things right. We need careful procedures and practices such as Institutional Review Boards or Ethics and Society Review (Bernstein et al., 2021) to ensure users' data can be used in appropriate and ethical ways (Bender et al., 2021), especially when it comes to protected information that is often manifested unconsciously by users in so-called "publicly observable" social interactions. It is necessary and essential to share data and models in social NLP to facilitate follow-up research; however, even if properly anonymized, data might contain clues to users' identity, and adversary can perform training data extraction attacks to recover personally identifiable information such as names and phone numbers by querying large pretrained language model (Carlini et al., 2020).

(6) Does Social NLP speak English? Most of today's research mainly focuses on 10 to 20 high-resource languages with a special focus on English, though there are thousands of languages and dialects with billions of speakers in the world. Language, dialect and the culture behind largely influences the comprehension of social NLP. For instance, Blodgett et al. (2016) found that existing language identification and dependency parsing tools on African-American Vernacular English text demonstrated very poor performances compared to on Standard English text. As NLP is now applied to everyday interaction globally, meaningful and impactful technologies will have to thoroughly model these social factors to avoid hegemonic approaches assuming all conversations follow Western culture and norms.

Acknowledgement I would like to thank Dirk Hovy for early discussion on related topics and the anonymous reviewers for their valuable feedback.

References

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Bernstein, Michael S., Margaret Levi, David Magnus, Betsy Rajala, Debra Satz, and Charla Waeiss. 2021. Esr: Ethics and society review of artificial intelligence research.

Blodgett, Su Lin, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Bowman, Samuel and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.

Brown, Penelope and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *CoRR*, abs/2012.07805.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Flek, Lucie. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838.

Grice, Herbert P. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Halliday, Michael Alexander Kirkwood and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Hovy, Dirk and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

Kiela, Douwe, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Lefcourt, Herbert M. 2001. *Humor: The psychology of living buoyantly*. Springer Science & Business Media.

Moravec, Hans. 1988. *Mind children: The future of robot and human intelligence*. Harvard University Press.

Nguyen, Dong, Laura Rosseel, and Jack Grieve. 2021. On learning and representing social meaning in nlp: a sociolinguistic perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 603–612.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis)contents: A survey of dataset development and use in machine learning research.

Pinker, Steven. 2003. *The language instinct: How the mind creates language*. Penguin UK.

Vincent, James. 2021. Openai's latest breakthrough is astonishingly powerful, but still fighting its flaws. In *The Verge*.

Wittgenstein, Ludwig. 2010. *Philosophical investigations*. John Wiley & Sons.

Crowdsourcing Relative Rankings of Multi-Word Expressions: Experts versus Non-Experts

David Alfter[1], Therese Lindström Tiedemann[2], Elena Volodina[1]

[1]University of Gothenburg
Språkbanken
Department of Swedish

[2]University of Helsinki
Department of Finnish, Finno-Ugrian
and Scandinavian Studies

david.alfter@gu.se
therese.lindstromtiedemann@helsinki.fi
elena.volodina@gu.se

Abstract In this study we investigate to which degree experts and non-experts agree on questions of difficulty in a crowdsourcing experiment. We ask non-experts (second language learners of Swedish) and two groups of experts (teachers of Swedish as a second/foreign language and CEFR experts) to rank multi-word expressions in a crowdsourcing experiment. We find that the resulting rankings by all the three tested groups correlate to a very high degree, which suggests that judgments produced in a comparative setting are not influenced by professional insights into Swedish as a second language.

1 Introduction

Many of the challenges in automatically driven solutions for language learning boil down to the lack of data and resources based on which we can develop language learning materials or train models. Resources like the English Vocabulary Profile (Capel, 2010, 2012; Cambridge University Press, 2015) are a luxury that cost a lot of time and resources to create, and for most languages such resources do not exist. Crowdsourcing has been suggested as one of the potential methods to overcome these challenges. Recently, a European network enet-Collect[1] (Lyding et al., 2018) has been initiated to stimulate synergies between language learning research and practice on the one hand, and crowdsourcing on the other. New initiatives have arisen as a result, e.g. using implicitly crowdsourced learner knowledge for language resource creation (Nicolas et al., 2020), crowdsourcing corpus cleaning (Kuhn et al., 2019), development of the Learning and Reading Assistant LARA (Habibi, 2019). However, there are many questions that need to be investigated and answered with regards to methodological issues arising from us-

[1]https://enetcollect.eurac.edu/

ing crowdsourcing as a method in/for Language Learning.

In this article, we raise some methodological questions about crowdsourcing in the context of second language (L2) learning material creation. To go back to the example of the English Vocabulary Profile – could we generate something similar for other languages without involving lexicographers and experts? For example, given a set of some unordered vocabulary items (e.g. phrases), how can we order them by difficulty and split them into groups appropriate for teaching at different levels of linguistic proficiency? Could a crowd help us in this scenario? Who can be "the crowd" in that case? How many answers are enough? How many contributors are needed? Are the results reliable? Parts of this article have been described in Alfter et al. (2020).

We focus on whether a crowd of non-expert crowdsourcers can be used to generate language learning materials and how the annotations by experts such as L2 Swedish teachers, assessors and researchers, i.e. people with formal training in teaching and assessing in Swedish, compare to the annotations by non-experts, by which we mean learners and speakers of L2 Swedish.[2] On a more general note, we investigate whether crowdsourcing as a method can be *reliably* applied to language learning resource building using a mixed crowd.

We use a selection of multi-word expressions (MWE) and ask experts (teachers, assessors etc) and non-experts (language learners) to arrange MWEs by difficulty. The crowdsourcing part of the experiment is designed in such a way that we test which *intuitions* people have about the relative difficulty of *understanding* word combinations. In this design, we do not expect our participants to know anything explicitly about language learning theories, instead relying on their intuitive comparative judgments as intuitive comparative judgments – including ranking items against each other – has been proven to be easier than assigning items to a category (e.g. a level of proficiency) (Lesterhuis et al., 2017). We hypothesize that given an unordered list of expressions, using crowdsourcing, we can derive a list ordered by difficulty that can be used in language teaching. We surmise that difficulty and proficiency are correlated, thus one might expect more difficult expressions to be learned at later stages of language development.

The theoretical notion of *L2 proficiency* is of special importance in connection to this study. Proficiency is a key concept in Second Language Acquisition (SLA) research. It is used to describe the language "knowledge, competence, or ability" (Bachman et al. (1990, p. 16), as cited in Carlsen (2012, p. 163)) of a learner on a conventionalized scale, one example being the 6-level scale adopted by the Common European Framework of Reference (CEFR) (Council of Europe, 2001, 2018). Conventionalized scales of proficiency levels are useful in educational and assessing contexts, e.g. which group to place a student into (Bachman et al., 2010) and in various social and political scenarios, e.g. whether an applicant can be granted citizenship (Forsberg Lundell, 2020). However, a straightforward division into levels is a tricky endeavor, since there is no consensus how to define a level and its corresponding competence(s) in concrete terms. SLA research is specific about viewing proficiency as a "coarse-grained, externally motivated" construct (Ortega, 2012, p. 134), where levels are always somewhat arbitrary (Council of Europe, 2001, p. 17) and proficiency should be seen as different to *L2 develop-*

[2] By L2 Swedish we mean Swedish as a second (third, fourth, …) language and as a foreign language

ment which is "an internally motivated trajectory of linguistic acquisition" (Ortega, 2012, p. 134).

For this reason, current approaches to proficiency advocate rather a scalar/interval approach as it is more powerful, realistic and nuanced (Ortega, 2012; Council of Europe, 2018; Paquot et al., 2020). The current experiment is proof of the usefulness of such an approach where rather than stating that certain vocabulary belongs to a certain level, we can instead state that some vocabulary items are perceived as easier or more difficult in comparison to each other and form a growing scale of items which are likely to be learned in that approximate order.

This article is structured as follows: we introduce related work in Section 2 and describe the data used for the experiment is Section 3. Sections 4 and 5 introduce Methodology and Experimental design. In Sections 6 and 7 we present our main results, analyze and discuss them. Section 8 concludes the article.

2 Related Work

Previously, several approaches have been used in identifying and ordering relevant vocabulary items for second language learning. A popular approach is to use reading material written by first language (L1) speakers such as newspapers to generate frequency-based word lists, e.g. the Kelly lists (Kilgarriff et al., 2014), the General Service List (West, 1953), the New GSL (Brezina and Gablasova, 2015). Such word lists tend to use frequency of occurrence in a corpus as the only criterion for deciding which items that should be taught first, and which ones should be introduced later, following the hypothesis that more frequent words would be easier (cf e.g. Eskildsen (2009) regarding usage-based approaches to L2 acquisition) and more important to know, while more rare words would be more difficult and less critical for communication in a target language. While such lists are useful, they also have drawbacks, especially in the context of second language learning. Indeed, L1 reading material is rarely adequate for language learner needs and lacks important vocabulary items (François et al., 2014, p. 3767).

In order to address the L2 learner needs, there has also been work on using L2 materials as a basis for word lists. One possible approach is to use graded textbooks as a starting point, as has been done in the CEFRLex project.[3] The motivation behind this approach is that textbooks generally target a specific proficiency group of language learners and have been carefully written with the needs of second language learners in mind. The project so far has resulted in the creation of six corpus-based language lists in six languages: FLELex for French (François et al., 2014), SVALex for Swedish (François et al., 2016), EFLLex for English (Dürlich and François, 2018), NT2Lex for Dutch (Tack et al., 2018) and ELELex for Spanish (François and De Cock, 2018). Each of these word lists not only contains the overall frequency but also the distribution of frequencies over the different CEFR levels. These projects have assumed that, in theory, the level at which a text is used in a language learning scenario can be used as an indication of a level at which vocabulary of that text can be assumed to be understood by learners and thus can be qualified as a learning target. In practice, however, this relationship is not as straightforward (e.g. Benigno and de Jong (2019)).

Another approach based on L2 material is to use graded learner essays. This

[3] https://cental.uclouvain.be/cefrlex/

has been done in projects such as the English Vocabulary Profile (EVP)[4] (Capel, 2010, 2012) and SweLLex (Volodina et al., 2016b). SweLLex belongs to the CEFRLex family, as it has been created with the methodology behind CEFRLex, but in contrast to other resources in the family, it is based on learner essays, more specifically the SweLL pilot corpus (Volodina et al., 2016a) of graded essays written by learners of Swedish. Both of these resources have also experimented with a threshold approach to assigning levels (Hawkins and Filipović, 2012; Alfter et al., 2016), i.e. taking as indicative level not simply the first occurrence but the first *significant* occurrence, i.e. the level at which a word or expression is used a certain number of times as defined by a threshold value. Deriving word lists from learner essays may prove more reliable as the amount of data increases (Pilán et al., 2016), and when the non-standard learner language has been effectively standardized (i.e. corrected) to the target language forms since automatic annotation is almost always trained on standard L1 materials (cf. Stemle et al., 2019). Both aspects, however, are non-trivial and very few languages enjoy the luxury of extensive corrected collections of learner-produced data.

Finally, one can consult L2 experts to rely on their judgments as to the difficulty of items. Expert judgment as a method has been widely applied in general linguistics as well as in second language oriented experiments and L2 resource creation (e.g. Spinner and Gass, 2019; Capel, 2010, 2012), although not without criticism. One of the potential stumbling blocks is the *subjective intuitive* nature of judgments, something which is claimed to be a major obstacle to reliable scientific conclusions; observations, i.e. language *production*, are regarded as a more reliable and desirable source of data (Bloomfield, 1935). However, Chomsky and Halle (1965) argue that judgments versus observations reflects the dichotomy between competence versus performance. In the end, expert judgments reflect experts' professional experience, and are based on evidence coming from *their* practices and theoretical assumptions about L2 teaching, and thus inevitably reflect personal interpretations of these. The challenge is, thus, to overcome the subjectivity of judgments without losing correctness of the final conclusions, so that the results can be used as a basis for assumptions about language learning paths and for scheduling learning materials in an optimal (although obviously never perfect) way. By *direct labeling* we mean that experts explicitly label each item with a CEFR level (A1-C2+). This method is also referred to as the *"Hey Sally"* method in Spinner and Gass (2019), indicating decision making based on consulting with other expert colleagues to either reduce or confirm the personal subjective bias.

Due to problems with the reliability of manual level assignment, some people have experimented with the number of experts and procedures that would be necessary to gain reliable objective results. Carlsen (2012) notes that the Norwegian L2 corpus project ASK (Tenfjord et al., 2006) used 10 CEFR assessors for their essays, who for the most part worked in groups of 5 so that each essay was marked by at least 5 assessors to get a reliable result. Similarly, Leńko-Szymańska (2015) used 2-4 raters for the level assignment of her subset of the international corpus of learner English (ICLE) (Granger et al., 2009) to reach agreement between the raters and Díez-Bedmar (2012) reported very low inter-rater reliability when using only 2 raters to assign CEFR-levels to Spanish university entrance exams. Furthermore, previous re-

[4] https://www.englishprofile.org/wordlists

search has shown that the background of the rater is of much importance (cf Díez-Bedmar (2012) for an overview), although the results have been mixed. Experienced raters have sometimes rated more strictly (Sweedler-Brown (1985) as cited in Díez-Bedmar (2012)) but in other studies they were more lenient (Weigle (1998) as cited in Díez-Bedmar (2012)). Whether the rater is a native speaker or not has also been seen to have an effect, in addition to gender, but once again the results were mixed. Díez-Bedmar (2012) also shows that how different rater backgrounds rate the proficiency has also depended on whether holistic or analytic scales were used.

The inherent order of teaching the items on the various vocabulary lists mentioned earlier, however, is not always obvious. Frequencies can be misleading, insufficient or sometimes idiosyncratic. Expert judgments might be perceived as less idiosyncratic but can be inaccessible due to the costs entailed in expert work. Crowdsourcing as a method of annotation could be worth exploring to address the above mentioned weaknesses.

To the best of our knowledge, crowdsourcing has not been extensively used for such ordering tasks. However, we surmise that it might be an alternative to the more heavily resource reliant methods. Crowdsourcing can take different forms. On the one hand, it can be quite explicit about the crowdsourcing aspect. In its original form, it would consist in the annotation of the same data by different annotators (Fort, 2016) or the collaborative creation and curation of resources such as Wikipedia (Stegbauer et al., 2009). Such forms generally rely on intrinsic motivation. However, if there is a lack of intrinsic motivation for whatever reasons, two different approaches have been taken, the first of which is paying people, and the second of which is making the task more fun by adding game-like elements (Chamberlain et al., 2013). The monetary aspect is expressed in platforms such as Amazon Mechanical Turk which pays participants to answer questions and/or solve tasks (Buhrmester et al., 2016). On the other hand, crowdsourcing can be more subtle, such as in Games With A Purpose (GWAPS) (Lafourcade et al., 2015). GWAPS are games or gamified platforms that serve a specific purpose which is not merely ludic.

There is research on creating language resources using crowdsourcing, some of which are: Zombilingo for syntactic annotation (Fort et al., 2014), Phrase detectives for co-reference annotation (Chamberlain et al., 2008) or JeuxDeMots for the creation of a lexico-semantic network (Lafourcade and Joubert, 2008). However little work has been done on the combination of crowdsourcing and language *learning*. Probably the most well-known approach on combining crowdsourcing and language learning was done by Duolingo (Garcia, 2013), although besides the stated goal of "translating the web while learning a language", it is not quite clear how the output is used. Recently, the use of implicit crowdsourcing techniques using language learners for the creation of language resources on par with expert-created content has also been explored (Nicolas et al., 2020).

A related field of work is crowdsourcing for education, of which the closest subaspect pertaining to this work is the creation of educational content. Initiatives include for example crowdsourced textbook generation (Solemon et al., 2013) or crowdsourcing video captioning correction by language learners to enhance learning (Culbertson et al., 2017). The interested reader is referred to Jiang et al. (2018) for an extensive review of current literature and practices.

3 Data

COCTAILL (Volodina et al., 2014) is a corpus of coursebooks for Swedish as a second language that we used as the basis for identification of candidate multi-word expressions (MWEs) for this experiment. COCTAILL contains texts and exercises aimed at adult learners of Swedish, and covers five CEFR levels: A1, A2, B1, B2, C1, where A1 is beginner level and C1 is advanced level (Council of Europe, 2001), with several coursebooks at each level (see Table 1). In the corpus, each chapter (lesson) in a coursebook has been assigned a level at which it is known to be used in an L2 teaching context. For example, suppose a textbook T contains 9 chapters and that practicing teachers are using chapters $T1$-$T4$ when teaching students aiming for the A1 level, and chapters $T5$-$T9$ aiming for the A2 level. All texts that are used in chapters $T1$-$T4$ are surmised to target A1 level knowledge, while texts that are used in chapters $T5$-$T9$ are assumed to target A2 knowledge, and so on. Further, all words that are used in the texts in chapters $T1$-$T4$ are labeled as potential target receptive vocabulary for the A1 level. All new vocabulary items that are used in texts in chapters $T5$-$T9$ (and that have not been used at previous levels) are labeled as potential target vocabulary at the A2 level, and so on. This approach allows us to generate useful vocabulary lists for both pedagogical and assessment use, as well as for automatic classifications of various kinds. However, generalizations about the levels at which vocabulary items should be targeted remains only an assumption that needs to be confirmed. Thus, the projected levels at the word level can serve as indications that certain items might be easier or harder, although we make no claims about the correctness of these projections.

Table 1 shows an overview of the corpus, detailing how many books targeting each CEFR level that are included, how many authors we rely on, as well as the number of chapters, texts, sentences and tokens.

COCTAILL is annotated automatically with the Sparv-pipeline[5] (Borin et al., 2016) for base forms, word classes, syntactic relations, word senses, MWEs and some other linguistic aspects. MWEs are identified on the basis of Saldo lexicon (Borin et al., 2013) entries, which means that only MWEs that are contained in Saldo will be recognized. As Saldo is under active development, the automatic pipeline will probably be able to identify more MWEs in the future. From the annotated version of COCTAILL, we have generated a new version of the SVALex list (François et al., 2016) based on senses, as Sparv has been updated to include a word sense disambiguation module since the creation of the original list. Word sense distinctions are based on Saldo senses.

An entry in the list consists of a combination of a base form with its word class (i.e. a lemgram), plus a word sense. Polysemous items have several distinct entries in the list and different frequency counts are associated with each of the sense entries. Each item contains its frequency distribution across different CEFR levels where it occurred and is associated with the lowest CEFR level of the texts in which it is observed. Starting from the list of 1351 MWEs in the list, two annotators classified them manually according to a custom typology (Lindström Tiedemann et al., In preparation).

For the experiment, we chose three different groups of MWEs based on this manual annotation, aiming to select a wide yet balanced variety of different types of expressions. This resulted in the selection

[5] https://spraakbanken.gu.se/sparv

CEFR level	#Textbooks	#Authors	#Chapters	#Texts	#Sentences	#Tokens
A1	4	10	37	101	1581	11132
A2	4	10	105	232	4217	37259
B1	4	12	83	345	6510	79402
B2	4	8	31	314	8527	101583
C1	2	2	22	115	5085	71991
Total	18	42*	278	1106	25920	301367

* 26 unique

Table 1: Statistics over COCTAILL per level

of the following three groups: (1) interjections, fixed expressions and idioms,[6] (2) verbal MWEs and (3) adverbial, adjectival and non-lexical MWEs. For the sake of conciseness and spatial limitations, we will refer to group 1 as "interjections", to group 2 as "verbs" and to group 3 as "adverbs". Figures 1, 2 and 3 show the number of occurrences per group per level in the resource based on the first round of annotation. From each of these three groups, we selected 60 expressions to be used in the experiment, with 12 items for each CEFR level, for a total of $3 * 60 = 180$ expressions. Expressions were de-contextualized in the sense that we did not provide any example sentences illustrating the use and context of the expression. While this decision may hinder the decision making process, it ensures that decisions are solely based on the expressions themselves, as opposed to syntactic complexity or other features that might be judged in a sentence.

Within each of the groups we prioritized items that had been classified and agreed upon by both annotators. We double-checked all items in the COCTAILL corpus to see that the *sense* we had listed was the one used in the corpus at the automatically assigned CEFR level; this step was necessary, as the automatic annotation of the corpus might not always identify the correct sense of a word or expression.

To make the experiment a learning experience and to make sure the level of difficulty was annotated in relation to a particular sense, we added definitions to all items. As far as possible we picked definitions from *Svensk ordbok* (svenska.se). When this was not possible, we used Saldo, Wiktionary, Lexin, or provided definitions of our own.

[6] We are aware of the difficulty of such distinctions. We tried to give strict definitions of fixed expressions and idioms as well as providing illustrative examples of both. However, comparisons of the annotations of the two annotators have shown that what annotator 1 classified as one of the categories could sometimes be annotated as one of the other categories by the other annotator which is why we decided to have these as a joint group for the experiment.

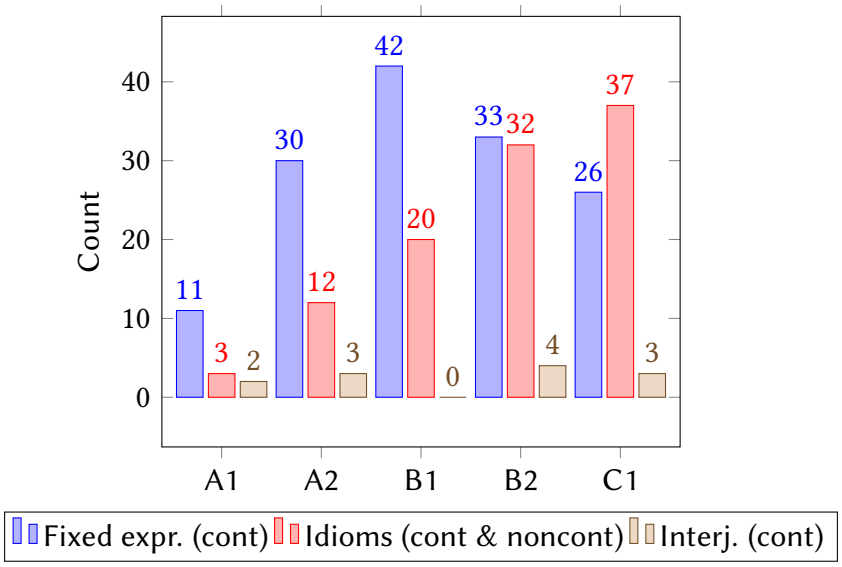


Figure 1: Group 1 in the crowdsourcing experiment

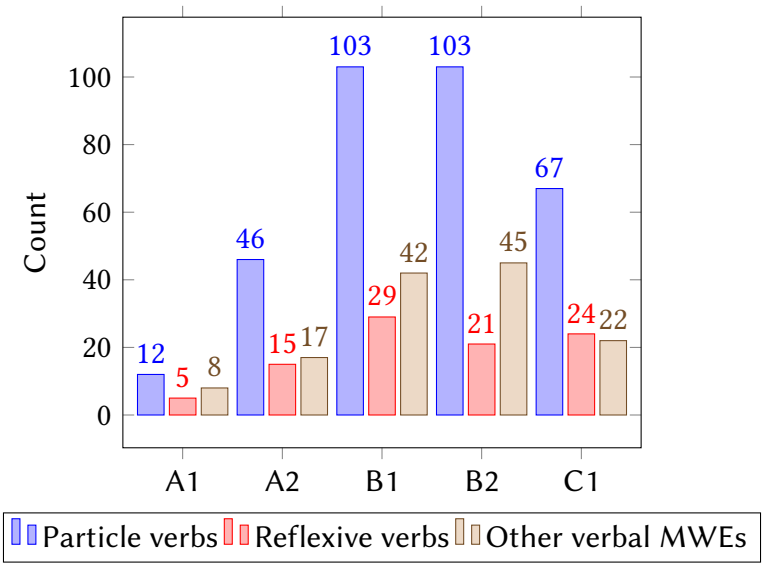


Figure 2: Group 2 in the crowdsourcing experiment

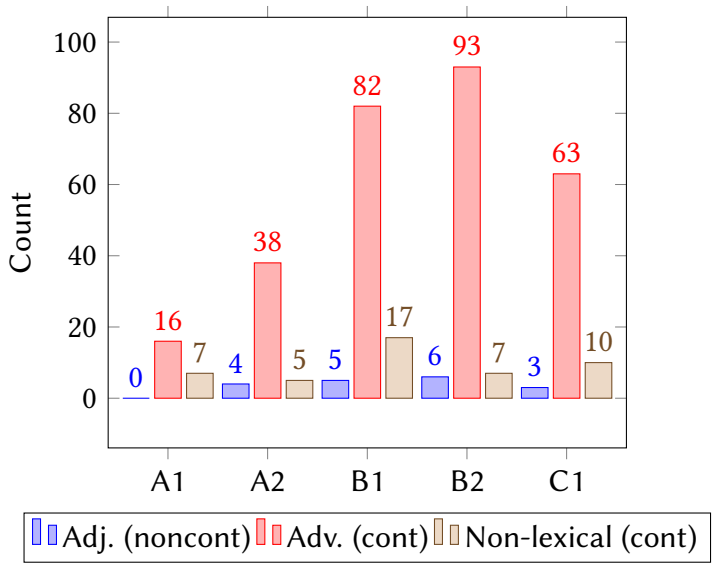


Figure 3: Group 3 in the crowdsourcing experiment

4 Methodology

Instead of having volunteers annotate each MWE with a target CEFR level (a task that requires in-depth knowledge of the CEFR), and following previous results showing that relative comparative judgments are easier than assigning items to a category (Lesterhuis et al., 2017), we opted to use best-worst scaling (Louviere et al., 2015) for the crowdsourcing task. The rationale is that language proficiency is a continuum rather than a set of discrete proficiency levels, although for practical reasons it is simplified to a set of discrete levels (Council of Europe, 2018, p. 34) (cf Section 1). Thus, using a relative ranking method may be more fruitful than trying to classify expressions into discrete classes; in addition, operating on a continuous scale allows for more sophisticated statistical measures to be used (Ortega, 2012, p. 131). Further, using best-worst scaling we get a maximum amount of information with a minimal amount of clicks from the crowdsourcers (Chrzan and Peitz, 2019). Finally, such a set up requires no knowledge of the CEFR, as participants rely on their intuition when judging expressions against each other.

In best-worst scaling, one is presented with a group of items to rank and asked to rank one of the items as the "best" or the "easiest" and one of the items as the "worst" or the "hardest". If one presents four items to the annotator to be ranked, having them choose both the easiest and the hardest out of the four expressions, it will result in 5 out of 6 possible relations.

To illustrate this further let us consider an example to show that four expressions give us six relations. Indeed, among four expressions, there exist six possible relations. Let us consider an example with expressions A, B, C and D, and let us assume that we want to know which of the expressions A, B, C and D that is the easiest and which is the hardest. This means that we thus have the following combinations of items between these expressions:

- *AB*
- *AC*
- *AD*
- *BC*

- BD
- CD

As the relations are symmetrical, we do not need to consider other combinations such as B A, as it is identical to A B; saying that B is easier than A implies that A is harder than B. With best-worst scaling, if one chooses B as the easiest expression and C as the hardest expression, we have knowledge of the following relations:

- $B < C$
- $B < A$
- $B < D$
- $C > A$
- $C > D$

The first point is self-explanatory: as we have stated that B is easiest and C is hardest, B must be easier than C. The other relations follow logically. As we have declared B as the easiest item, B must be easier than any of the other items (points 2 and 3). As we have declared C to be the hardest item, it must be harder than all other items (points 4 and 5). The only relation that we do not have information about is the relation between items A and D. However, this relation will be covered by subsequent tasks in which A and/or D occur.

In order to cover all possible combinations using best-worst scaling, we have chosen a redundancy-reducing combinatorial algorithm to calculate the minimum amount of combinations of four items needed to cover all relations in such a way as to minimize redundancy, i.e. repeating items that have already been encountered, based on Čibej et al. (In preparation).

With four items per task and 60 expressions there are 1,770 possible relations and 487,635 possible combinations. Using the redundancy-reducing combinatorial algorithm, this means that we need to have 326 tasks. Of the 1770 relations,

- 1362 (77%) are non-repetitive
- 33 with 1 relation known
- 50 with 2 relations known
- 12 with 3 relations known
- 3 with 4 relations known
- 1 with 5 relations known

Thus 77% of the relations are covered by non-repetitive combinations, while 23% of the relations are covered by partially repetitive combinations.

Finally, using best-worst scaling leads to a decrease in effort spent on the task. If one were to rank four items out of four in relation to each other, one would need at least four clicks, while best-worst scaling requires (a minimum of) two clicks, reducing the workload by half.

5 Experimental Setup

One of the aims of this study is to test how one's background influences the outcome of a crowdsourcing experiment. To take a step towards that aim, we experiment with two different ways of ranking MWEs according to difficulty.

1. Intuition-based (implicit) labeling, i.e. crowdsourcing: We ask a heterogeneous group of L2 speakers of Swedish (non-experts) as well as experts (L2 Swedish professionals e.g. teachers, researchers) to rank items by taking part in a crowdsourcing experiment where we subdivide the expert group into a general L2 professional group and a group of CEFR-experts:

- Non-experts: L2 speakers of Swedish at intermediate level (B1) or above (according to self-assessment)
- Experts – L2 Professionals: Teachers, assessors and/or researchers of Swedish as a second language (referred to as L2 professionals)
- Experts – CEFR experts: A separate subgroup of L2 professionals who use CEFR in their L2 Swedish practices

2. Expert judgment-based (explicit) labeling: We ask a small group of CEFR experts (teachers/researchers/assessors) to label MWE items manually for the levels at which they expect L2 learners to understand them. This annotation task is formulated in *levels* rather than *relative ordering* to resemble a real-life annotation scenario as much as possible where experts would be involved – which, however, entails some difficulties in comparison of the results.

5.1 Practicalities

Figure 4 illustrates the steps necessary to take part in the experiment. In the first step of the experiment, to comply with the GDPR (EU Commission, 2016) we asked our participants for consent to use their background information for this research and to send out gift certificates.[7] At the same time, we collected information about the linguistic background as well as some other demographic variables as illustrated in Table 5.4.

After filling out the consent form, participants were provided with guidelines and links for the crowdsourcing part of the experiment in the form of an automated email sent to the email address specified in the consent form.[8] The guidelines were intentionally provided only in Swedish as a "selection" principle to exclude L2 speakers of lower proficiency levels.

In the second step, participants were asked to create an account on the crowdsourcing platform, with the explicit instruction to use the same email address as provided in the consent form so that we could link their background information to the crowdsourcing results. Email addresses were solely collected for this purpose and were discarded after this linking step was performed.

As a final step, participants were asked to participate in the projects proper. Each crowdsourcer was expected to complete at least 84 items out of 326 in each of the three projects, which amounted to a total time of about 30-45 minutes per project. Participants who completed at least 84 tasks per project were sent a gift (step 3 in Figure 4).[9]

To reach the crowdsourcing population, we published announcements via e-mail, social networks, and through professional and private networks. For CEFR experts, we listed requirements with regards to their qualifications and recruited three experts on the basis of this.

We left a calendar month for the crowdsourcing experiment from the date of the first announcement, with periodic re-

[7] Expert form (Swedish only): `https://spraakbanken.gu.se/larkalabb/mwe-cs-annotation-teacher`
Non-expert form (Swedish only): `https://spraakbanken.gu.se/larkalabb/mwe-cs-annotation-crowd`

[8] Guidelines (in Swedish): `https://docs.google.com/document/d/1E700mnqaZ15cHr_3gXMvg0d4onm2ncMf36t-KUQuRrw/`

[9] In the later stages of the experiment when it was not possible to contribute 84 tasks in one or more projects, we relaxed the constraints for gift eligibility to ≈ 240 tasks in total.

Figure 4: Practical steps for participants in the crowdsourcing experiment

minders to recruit broader participation. All crowdsourcers that met our requirements of the minimal contributions, were sent small gifts. Experts were paid by the hour.

5.2 Implementation

For the crowdsourcing experiment, we set up nine projects for the three different participant backgrounds. All projects were implemented in pyBossa, an open-source customizable framework for crowdsourcing tasks developed by SciFabric.[10] For each of our three target groups (Non-experts = L2 speakers, L2 Professionals = L2 teachers, researchers; CEFR Experts = L2 teachers, researchers, assessors with CEFR experience) we prepared three projects consisting of three sets of different MWE-types (3 participant groups x 3 projects = 9 crowdsourcing projects). In addition, we set up a tenth crowdsourcing experiment for people who did not conform to any of the three target groups or for people who wanted to see how the projects work.[11]

For each of the projects, we arranged the 60 selected items per MWE group in such a way that the crowd could vote on their relative difficulty. Figure 5 shows the graphical user interface we designed for this task, based on Čibej et al. (In preparation). In the user interface, crowdsourcers were shown four MWEs and were asked to indicate which expression they found the easiest and the hardest to understand by using the buttons on the left and the right of the expressions, based on their own intuition. In addition, one could click on any of the four expressions to be shown a definition in case one was not sure about the meaning of an expression. The interface also showed a pyBossa-internal ID number, the number of tasks that had been completed by the crowdsourcer, the number of total tasks (326 for each project) and the expected number of tasks that each crowdsourcer should finish (84 for each project, except for the "CEFR experts" who were expected to complete all 326 tasks). Finally, we also included a link to a feedback form where crowdsourcers could indicate their reasoning about assigning the labels for easiest and hardest, or any other feedback they may wish to provide.

As additional safe-guards, we implemented checks for user errors for the following cases:

1. No value selected
2. Only one column is selected
3. Same value in both columns

As we wanted to collect the easiest and the hardest expression among a set of four expressions, it was disallowed not to provide any value (point 1), to only choose either an easiest expression or a hardest expression but not both (point 2) or to select the same expression as both the easiest and the hardest (point 3). Furthermore, as we wanted to maximize user interaction, we took care to make sure that the platform was functional and usable not only on desktop PCs but also on smaller screens such as smartphones. By doing so, people could use their smartphones wherever they were and whenever they had a minute

[10] https://pybossa.com/

[11] Test-project (in Swedish): `https://ws.spraakbanken.gu.se/ws/tools/crowd-tasking/project/l2p_mwe_group2_other/`

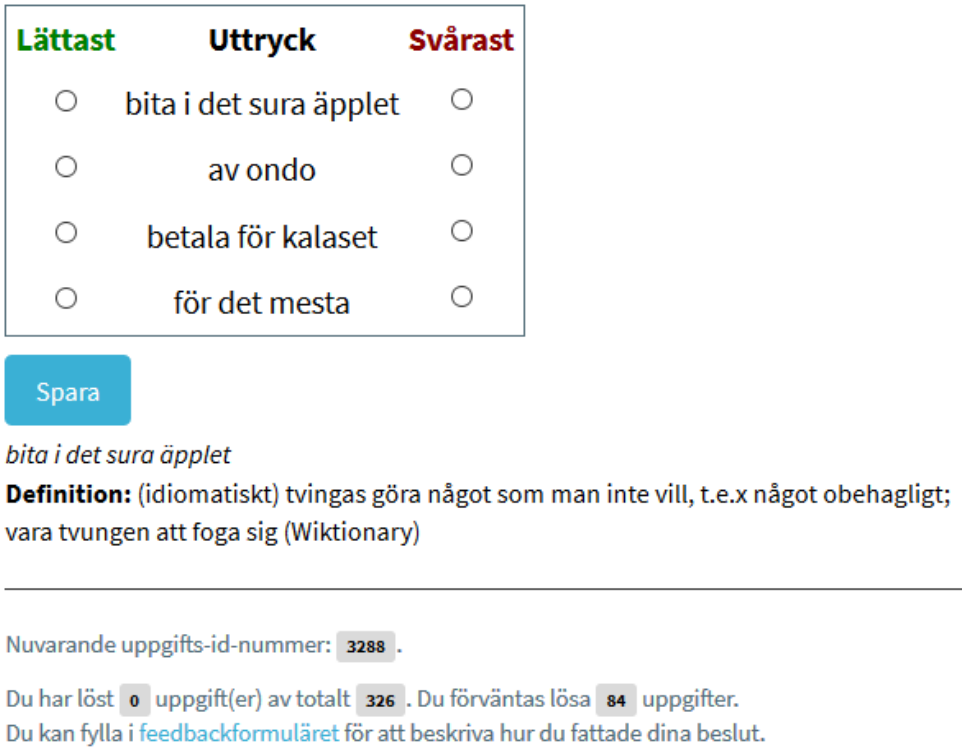


Figure 5: Example of an MWE ranking task in pyBossa (lättast = easiest, svårast = most difficult, uttryck = expression; spara = save)

to continue working on the tasks. Other considerations concerned the placement of the "easiest" and "hardest" columns, color schemes, and the ease of use on a smart phone. After registration in pyBossa participants could log in and continue from where they left off at any time suitable to them and on any platform (smartphone, tablet, computer). As to the number of votes per task, i.e. how many different answers were needed per task for a task to be considered complete, we set the number to 5 for L2 speakers and to 3 for L2 professionals and CEFR experts. These numbers were picked based on the estimated number of participants in the various groups. This meant that each single task in the project would have 5 respectively 3 answers (i.e. judgments about the easiest and the hardest expression) by different annotators.

We assigned the following scores to expressions: 1 for the expression that was rated as the easiest, 3 for the expression that was rated as the hardest and 2 for the two unrated expressions.

5.3 Experimental design

Figure 6 shows an overview of the experimental design. In the experiment, we wanted so see whether non-experts and experts agree with each other about the relative difficulty of multiword expressions in the crowdsourcing experiment. We further subdivided the expert group into two to be able to compare experts' indirect judgement (crowdsourcing) to their direct (explicit) labeling, but this was only done with the small subgroup of CEFR experts who we therefore had to make sure were all well familiar with CEFR in connection with their work. Finally, we wanted to check whether individual explicit labels by the CEFR experts coincided with the group re-

sults from their implicit crowdsourcing experiment.

As indicated above we also asked our three CEFR experts to perform a direct labeling task. This meant that we asked them to go through all of the selected MWEs in a spreadsheet and decide at which CEFR level these MWEs could be expected to be understood. All three CEFR experts were asked to do the crowdsourcing experiment in pyBossa first and were only given access to the spreadsheet for direct labeling after they had completed that, to make their crowdsourcing experience as similar to that of the rest as possible. However, unlike the other participants the CEFR experts were asked to rank all items in all the three pyBossa projects ($3 * 326$). In the direct annotation experiment, they were asked to pick one level from a drop-down menu with A1, A2, B1, B2, C1, C2 or above for each item in a spreadsheet with all 180 MWEs.

5.4 Demographic information

To better understand whether and/or how the intuitions and judgments were influenced by the background of the participants, we collected information about our participants in a separate form (personal metadata). Since L2 Swedish is widely spread in Sweden and Finland, these two countries were our primary targets. However, we used social media and our personal professional networks to spread information about the experiment, which also encouraged participation from other countries. Out of 79 consent registrations in total, 50 crowdsourcers participated in the experiment, which constitutes a drop-out rate of 37%. Upon completing the crowdsourcing experiment, we could see the following participant characteristics (Table 2):

We attracted 27 L2 non-experts (L2 speakers) and 23 L2 experts, including the three CEFR experts. Sweden and Finland contributed with 22 participants each (at 44%). The first language of the contributors is dominated by Finnish (30%), but other first languages are also represented, including Swedish (20%), German (12%), Russian (8%), Spanish (4%), Arabic (4%), Hungarian (4%) and others. The population is well-educated having either a pre-doctoral university degree (60%) or a doctoral degree (36%). L2 speakers provided self-assessed levels of Swedish as B1 or above in 96% of the cases, with one outlier at the A1 level. 65% of the L2 experts have 10 years or more of experience of teaching or assessing Swedish as a second language. The age characteristics show that we attracted a rather "mature" population (78%), whereas people of 30 years and younger are less represented (22%). The gender representation is rather unbalanced (66% women versus 28% men with 6% who preferred not to answer that question), which can be due to a recruitment bias or – potentially – reflect gender representation within the areas of language learning and teaching. All in all, we have participants of various background profiles, which represents the target group for the intended output of the research.

The three CEFR experts recruited come from Finland since Finland appears to use CEFR more extensively in the teaching and assessment of L2 Swedish than Sweden does. All CEFR experts have Finnish as their L1 and represent: one L2 Swedish teacher, one L2 Swedish researcher (PhD) and one L2 Swedish assessor (PhD).

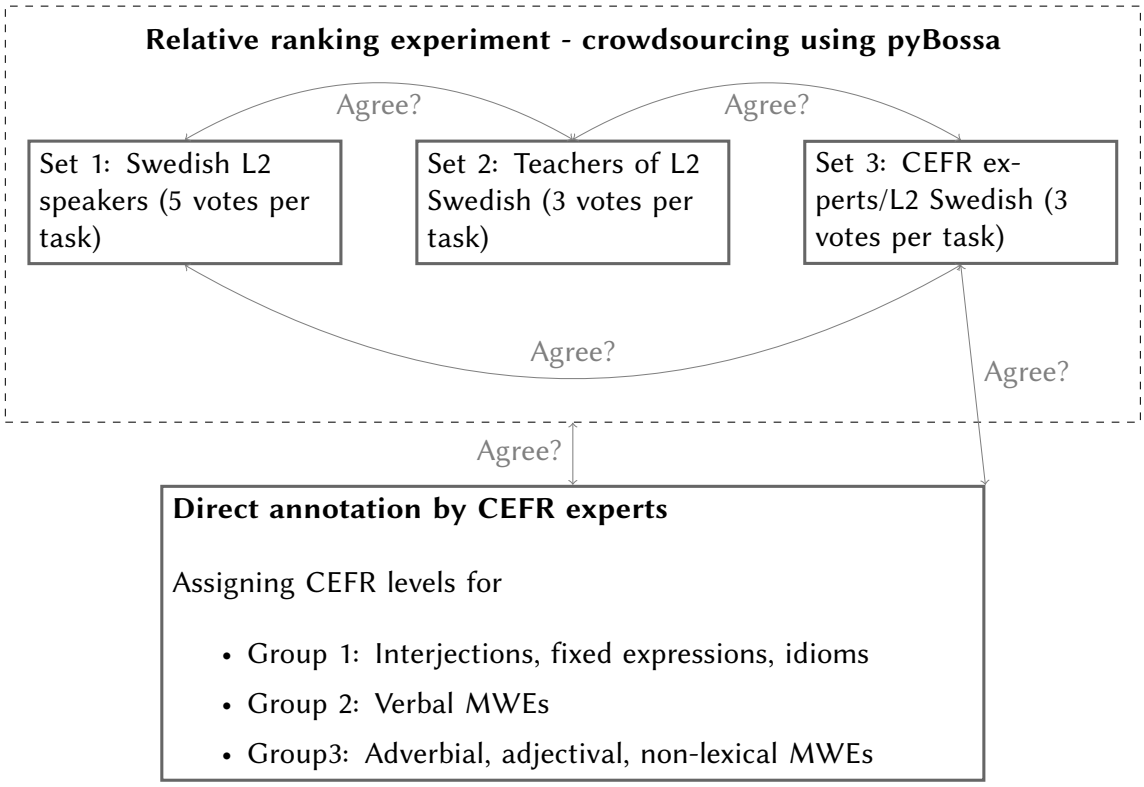


Figure 6: Overview of the experimental design

Table 2: Demographic variables

	Profiles	L2 speakers	L2 experts	Finland L2 speakers	Finland L2 experts	Sweden L2 speakers	Sweden L2 experts	Other L2 speakers	Other L2 experts
Total	50	27	23	9	13	13	9	5	1
Gender									
Female	33	15	18	7	10	7	8	1	-
Male	14	9	5	2	3	4	1	3	1
Other	3	3	-	-	-	2	-	1	-
Age									
16-20	5	5	-	3	-	2	-	-	-
21-30	6	2	4	2	4	-	-	-	-
31-40	15	10	5	4	3	3	2	3	-
41+	24	10	14	-	6	8	7	2	1
Education									
High school	2	2	-	-	-	2	-	-	-
University	30	16	14	9	8	5	6	2	-
PhD	18	9	9	-	5	6	3	3	1
Mother tongue									
Arabic	2	2	-	-	-	2	-	-	-
Dutch	1	1	-	-	-	-	-	1	-
English	1	1	-	-	-	1	-	-	-
Finnish	17	8	9	8	9	-	-	-	-
German	6	5	1	1	-	3	-	1	1
Hungarian	2	2	-	-	-	1	-	1	-
Luxembourgish	1	1	-	-	-	1	-	-	-
Norwegian	1	1	-	-	-	1	-	-	-
Russian	4	3	1	-	-	3	1	-	-

Table 2: Demographic variables continued

	Profiles	L2 speakers	L2 experts	Finland L2 speakers	Finland L2 experts	Sweden L2 speakers	Sweden L2 experts	Other L2 speakers	Other L2 experts
Serbian	2	-	2	-	1	-	1	-	-
Slovenian	1	1	-	-	-	-	-	1	-
Spanish	2	2	-	-	-	1	-	1	-
Swedish	10	-	10	-	3	-	7	-	-
L2 Swedish level									
A1	1	1	-	-	-	-	-	1	-
A2	-	-	-	-	-	-	-	-	-
B1	4	4	-	2	-	-	-	2	-
B2	7	7	-	2	-	4	-	1	-
C1	9	9	-	4	-	4	-	1	-
C2	6	6	-	1	-	5	-	-	-
Teaching years									
1-9	8	-	8	-	6	-	2	-	-
10-19	7	-	7	-	3	-	3	-	1
20+	4	-	4	-	1	-	3	-	-
other	4	-	4	-	4	-	1	-	-

5.5 Evaluation methodology

We use two modes of annotating items for their difficulty: crowdsourcing by non-experts and experts, and direct annotation by experts. Since direct expert annotation is a rather traditional approach, we use traditional ways of evaluating it, relying on metrics such as agreement and Spearman rank correlation. However, crowdsourcing is a new approach for this type of tasks, thus we explain how we evaluate and compare the results of the crowdsourcing experiment below. The results are presented in Section 6.

For evaluation of the crowdsourcing, we project each expression onto a linear scale. The scale ranges from 1 to 3, with a minimum value of 1 if an expression was always classified as the easiest out of a set of four possible expressions and a maximum value of 3 if it was always classified as the most difficult out of a set of four possible expressions by all annotators. Otherwise, the expression *exp* is assigned the score $s(exp)$ as the mean of all assigned scores x according to the formula

$$s(exp) = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

With x_i being the i-th score assigned to exp and n being the total number of scores assigned to exp. The limits of 1 and 3 are predetermined by our way of measuring "the easiest" and "the hardest" expression with best-worst scaling (cf Section 4).

We sort the data according to the reverse order of $s(exp)$ and assign sequential ranks from 1 to 60 to the resulting ordering.

6 Results and Analysis

In this section we present the results from the different experiments. First, we look into the results from the crowdsourcing experiment. After this we look into the results from the direct annotation (expert labeling) which we also compare to the rankings which this group of experts did in pyBossa. Finally, we also investigate how the number of votes influences the results and how much time is needed for the crowdsourcing experiment as opposed to direct expert annotations.

6.1 Linear scale

The experiments generated rich data for analysis. In this section we look at the results of the study from a quantitative point of view. For the purpose of comparison, we projected results of crowd-votings to linear scales based on the fact that each vote in a crowdsourcing task assigns scores to the items: either 1 ("easiest"), 3 ("most difficult"), or 2 for each of the two items in-between. Based on the numerical values, all items are listed in the order of their scores corresponding to the perceived degree of difficulty.

Based on that principle, we obtained one linear scale per participant group and one representing the whole population of crowdsourcers (mixed background rankings).

Table 3 shows the Spearman rank correlation coefficient between the three sets of MWEs and the three groups of participants. Spearman rank correlation coefficient has a range from -1 to +1 where -1 indicates a perfect negative correlation; zero indicates no correlation; and +1 indicates perfect positive correlation.

As can be gathered from Table 3, the highest correlations can be found between non-experts (here meaning L2 speakers/learners) and the general group of "L2 professionals" (including teachers, assessors, researchers) across all of the three MWE groups, while the correlations between non-experts (L2 speakers) and "CEFR-experts" (i.e. the subgroup of three

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
L2 speakers-L2 professionals	0.9509	0.9282	0.9203
L2 speakers-CEFR experts	0.9333	0.8115	0.8370
L2 professionals-CEFR experts	0.9386	0.8495	0.8579

Table 3: Agreement between voter groups in the crowdsourcing experiment

L2 professionals) are the lowest among all the three MWE groups. We can thus say that non-experts (L2 speakers) and experts (L2 professionals) in our experiment agree very well on the relative difficulty of MWEs, followed by L2 professionals and CEFR experts, while L2 speakers and CEFR experts tend to agree to a lesser extent. Despite these marginal fluctuations, we can see strong correlations between all of the tested target groups across all the three sets of tested MWEs. This indicates that intuitions about the difficulty of MWEs are more or less shared across all tested groups, despite the differences in background and professional competence. It seems that we can confirm that non-experts – that is, L2 speakers lacking expertise and competence in a subject (e.g. language assessment) – can be seen as on par with experts for tasks requiring high competence, something that has also been shown in approaches in citizen science (Kullenberg and Kasperowski, 2016).

To get an insight into how well individuals can agree on crowdsourcing tasks we looked at the three CEFR experts in our experiment who completed the full sets of tasks in all of the three pyBossa projects. Table 4 shows the Spearman rank correlation based on their individual linear scales calculated from the crowdsourced data. As can be seen from Table 4, annotators 1 and 2 tend to agree the most, while annotators 1 and 3 tend to agree the least, with annotators 2 and 3 falling in-between. This might be a result of their different backgrounds and how often they use CEFR explicitly. The more voters we have, the less bias there is in the resulting data (e.g. Snow et al., 2008).

6.2 Expert labeling

If we look closer at the simple and extended percentage agreement between the CEFR expert annotators in the explicit (interchangeably called 'direct') labeling experiment, we can see that agreement is generally quite low for simple agreement (Tolerance 0 in Table 5). With a tolerance of zero, one counts exact agreement between the annotators (e.g. the same item has been assigned to the same CEFR level). However, if one relaxes the tolerance level to 1 (extended percentage agreement), meaning that positive agreement also includes cases where annotators differed by only one level (e.g. one annotator said the item was A2 while another annotator said the item was B1), we can see that agreement drastically improves, as illustrated in Table 5.

In general, this gives us a picture that expert judgments are not ideal and that reaching an exact agreement between them is possibly an unattainable target, which also confirms the results from essay evaluation according to the CEFR-scale as presented in e.g. Díez-Bedmar (2012). Given that direct labeling is a subjective and cognitively challenging task, more opinions than one are required (cf Snow et al., 2008; Carlsen, 2012). The MWEs in the experiments are de-contextualized which might further complicate decisions.

	Gr.1 (interj.)	Gr.2 (verbs)	Gr.3 (adv.)
CEFR experts 1 and 2	0.8130	0.8581	0.7735
CEFR experts 1 and 3	0.7733	0.5788	0.6988
CEFR experts 2 and 3	0.7964	0.6236	0.7026

Table 4: Inter-annotator agreement for CEFR experts in the crowdsourcing experiment calculated with Spearman rank correlation coefficient

	Group 1 (interjections)	Group 2 (verbs)	Group 3 (adverbs)
Tolerance 0	15.00	21.70	13.30
Tolerance 1	61.70	58.30	65.00

Table 5: Agreement between CEFR experts in a direct labeling experiment in percent

This speaks in favor of assuming tolerance level 1 since the assigned levels describe a continuum of proficiency rather than strict categories (Council of Europe, 2018, p. 34). A hypothesis in connection to this is that disagreement outside tolerance 1 may indicate items that are on the periphery of the lower CEFR level, while items within tolerance 1 constitute the core vocabulary on the lower level. This is something to be explored in future research.

Results of agreement between the explicit ranking of each individual expert and their own individual implicit judgment from the crowdsourcing experiment based on a comparison of the linear scales show mixed results (Table 6).

Expert 1 is very consistent in both annotation methods, and all annotators seem to agree with themselves most for MWE group 1, while other agreements are lower. This could indicate that expert 1 is the one with the most experience with working with CEFR-levels. The inconsistency of the results for the same expert indicates that the expert reasons differently when using different methods, and that the way of reasoning influences the results. It has been previously shown that explicit scoring is more subjective and cognitively demanding than assessing by comparing two samples to each other (Lesterhuis et al., 2017), which also seems to be confirmed in this experiment. This indicates that we should not compare the two types of annotation and that expert judgment can only give reliable annotation if a reasonably large number of experts is used to counter-balance a potential subjective bias (cf. Snow et al. (2008)). How large a number constitutes a "reasonable amount" is still an open question.

6.3 Number of votes

Aker et al. (2012) found that using one set of non-expert results (results from different annotators) outperformed using one single non-expert's results, as the diversity of the crowd might cancel a high bias present in a single annotator. In order to see how the number of votes influences the results, we randomly selected votes for the sample sizes 1, 2 and 3 (for the non-expert crowd, for which we collected 5 votes) and derived the linear scales, for each group separately as well as a randomly sampled mixed version over all three groups ('Mixed' in Tables 7 and 8). We then compared the linear scales of the different sample sizes to the

	Group 1 (interjections)	Group 2 (verbs)	Group 3 (adverbs)
Expert 1	0.9095	0.9280	0.8935
Expert 2	0.8483	0.6147	0.7299
Expert 3	0.8010	0.5248	0.5540

Table 6: Spearman rank correlation coefficients for intra-annotator agreement between implicit and explicit modes of annotation

linear scale derived from the full set votes (3 for experts and 5 for non-experts; for the mixed group we calculated the target linear scale from a random sample of three votes from both experts and non-experts), meaning that we compare for example the linear scale for non-experts derived from a single vote versus the linear scale for non-experts derived from 5 votes; the linear scale for non-experts derived from two votes versus the linear scale for non-experts derived from 5 votes; or the linear scale derived from randomly sampling two votes from both experts and non-experts versus the linear scale derived from randomly sampling three votes over all groups.

In order to quantify the differences between the scales, we used the *out-of-place metric* m_{oop} (Cavnar et al., 1994). This is a straightforward metric that measures the difference between two ranked lists and quantifies the difference. The reason for choosing this metric over rank correlation measures is that Spearman's correlation coefficient was very high and had similar values across all comparisons (see Table 7). While a high correlation is a positive result in itself, it does not allow for a detailed analysis. We surmise that using m_{oop} may give a more tangible result. It is formalized as shown in (2)

$$m_{oop} = \sum_{i=1}^{n} (|r(x_i, l_1) - r(x_i, l_2)|) \quad (2)$$

with n being the number of items in the lists (the lists to be compared are of the same length in our case), x_i being the i-th item, $r(x_i, l_1)$ being the rank of x_i in the first list and $r(x_i, l_2)$ being the rank of x_i in the second list. To illustrate this, let us consider two lists l_1 and l_2 both containing the expression A, B, C and D, but at different ranks. Figure 7 shows a hypothetical scenario. In order to obtain m_{oop}, one first calculates the difference in ranks between the expressions, then sums up the differences. Thus, in this example, we would have $m_{oop} = 1 + 2 + 0 + 3 = 6$. We also calculate how many items are at the exact same rank in both lists (out of 60 total).

We find that each of the sub-sampled lists compared to the full-vote list yields high Spearman rank correlation coefficients, with Spearman's ρ varying from $\rho = 0.941, p = 5^{-29}$ to $\rho = 0.997, p = 3^{-66}$. As can be gathered from Table 7, group 1 (interjections) shows the least amount of divergence among all three MWE groups, but also among the different crowds. Further, it can be observed that sampling over all three crowd groups produces more stable results than within-group sampling.

A more qualitative analysis reveals that for group 1 (interjections etc.) for non-experts with one vote, the hardest and easiest item is the same as with five votes, whereas with two votes, the two easiest and the three hardest are the same as with five votes. For CEFR experts with one vote, the three easiest items are the same as with three votes, whereas with two votes, the two hardest items are also the same as with five votes. For L2 professionals with one

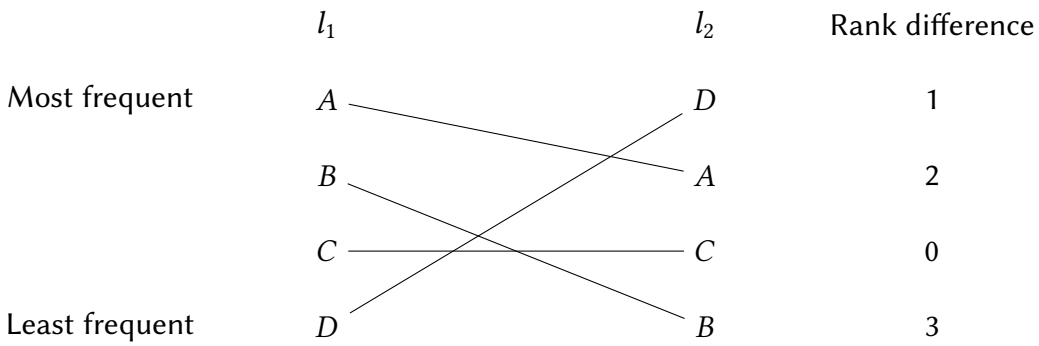


Figure 7: Out-of-place metric illustration

MWE group	Crowd	Sample size	m_{oop}	ρ	Same rank
Interj.	L2 sp.	1	150	0.98	8
		2	112	0.98	15
		3	102	0.99	16
	L2 prof.	1	160	0.97	16
		2	82	0.99	15
	CEFR exp.	1	114	0.98	18
		2	80	0.99	18
	Mixed	1	114	0.98	14
		2	78	0.99	23
Verbs	L2 sp.	1	256	0.94	6
		2	172	0.97	6
		3	114	0.98	18
	L2 prof.	1	196	0.97	8
		2	90	0.99	18
	CEFR exp.	1	200	0.96	7
		2	120	0.98	14
	Mixed	1	138	0.98	11
		2	70	0.99	26
Adverbs	L2 sp.	1	254	0.95	4
		2	154	0.98	12
		3	110	0.98	15
	L2 prof.	1	244	0.94	8
		2	132	0.98	14
	CEFR exp.	1	126	0.98	14
		2	106	0.99	13
	Mixed	1	128	0.98	14
		2	54	0.99	25

Table 7: Out-of-place calculations, Spearman's ρ and same rank number for different numbers of votes

MWE group	Crowd	Sample size	d					
			0	1	2	3	4	5
Interj.	L2 sp.	1	8	21	34	45	49	55
		2	15	30	43	48	51	56
		3	16	34	44	51	56	58
	L2 prof.	1	16	30	38	43	47	52
		2	15	38	53	56	58	59
	CEFR exp.	1	18	35	42	46	53	57
		2	18	42	49	55	57	59
	Mixed	1	14	30	44	49	55	57
		2	23	37	49	54	59	60
Verbs	L2 sp.	1	6	10	27	36	42	45
		2	6	26	37	42	48	52
		3	18	34	43	50	54	55
	L2 prof.	1	8	17	24	33	43	50
		2	18	37	47	54	56	58
	CEFR exp.	1	7	20	32	35	44	49
		2	14	26	38	50	56	57
	Mixed	1	11	22	36	45	52	57
		2	26	44	48	55	58	59
Adverbs	L2 sp.	1	4	16	27	31	37	40
		2	12	26	33	41	50	53
		3	15	36	46	53	54	55
	L2 prof.	1	8	17	29	36	41	46
		2	11	30	41	48	51	55
	CEFR exp.	1	14	31	44	48	51	54
		2	13	33	44	51	56	58
	Mixed	1	14	32	44	49	49	54
		2	25	48	54	59	60	60

Table 8: Effect of different d values

vote, the easiest item is the same as with three votes whereas with two votes, also the two hardest items are the same as with three votes. However, many of the rank differences are small, i.e. the two hardest items for group 1 (interjections etc) for L2 professionals with one vote are the reverse order of two and three votes. If one were to start from a truly unlabeled set of items without indications of level, or the number of different levels present in the data, one can only rely on relative ranks. These results indicate a certain stability when it comes to the extremes of the scale, i.e. which items are easiest and which items are hardest.

In order to account for small differences in ranks, we also compute how many items are "at the same rank" when counting as the same rank items within a difference of d, with d varying from 1 to 5 (n.b. $d = 0$ is equivalent to the same rank, column 'Same' in Table 7; cf 'Tolerance'). If we take as an example the ranking in Figure 7, at $d = 1$, one would count as being of equal rank the item A (in addition to item C), as the rank difference is 1. At $d = 3$, one would also consider as being of equal rank items B and D, as they are below or equal to 3. Table 8 shows the results; we repeat $d = 0$ for comparison purposes.

It can be said that the lists derived from a sub-sample of votes are different from the lists derived from all votes. However, when relaxing the notion of "equivalence" as has been done by varying d, one can see that the difference is not as big as one might think at first. At $d = 2$, which means deviations of two ranks (out of 60) or less are counted as equal, around 84% of the lists are "equal" to the lists derived from full votes for the aggregated versions (82% for interjections, 80% for verbs and 90% for adverbs). Again, it can be observed that sampling over the whole crowd produces more stable results than sampling within a group. It can further be observed that the aggregated votes tend to be on par with expert judgments, if not surpassing them.

6.4 Time investment

Table 9 shows the average time taken per crowd background and MWE group. Despite the presence of outliers in the non-expert crowd data, crowdsourcing in a best-worst scaling scenario takes on average 30-40 seconds per task. To rank 60 items presented through 326 tasks with one vote would claim ≈ 2,5-3 hours. Rankings do not seem to change drastically after the first three votes are collected, so the minimal time investment for 3 votes are estimated to approximately 8-9 hours for one project.

Table 10 shows the comparison between observed times in the crowdsourcing project and reported times for direct annotation by the CEFR experts. It should be noted that for expert direct annotation, the times indicated in Table 10 are approximated by dividing the reported time needed to finish all three lists by three. It should also be borne in mind that experts went through all 326 tasks per project. It can be observed that direct expert annotation claims 15-90 minutes per project. This is at least five times as fast as the crowdsourcing experiment.

However, reliability and consistency of a (direct) labeling depend to a larger extent upon what kind of ranking scale annotators are offered and what their backgrounds are, and the effects are difficult to account for (cf O'Muircheartaigh et al., 1995). It is easy to fall victim to a flawed design, inexperienced annotators or face problems hiring annotators, and the cognitive load of such an exercise is higher than in a crowdsourcing set-up (e.g. Lesterhuis et al., 2017).

The time required to complete such a

	Group 1	min	max	Group 2	min	max	Group 3	min	max
L2 speakers	36	3	164	38	6	260	44	3	227
L2 prof.	41	13	43	26	14	44	24	14	44
CEFR exp.	32	28	39	34	23	41	36	21	60
Average	36			32			34		

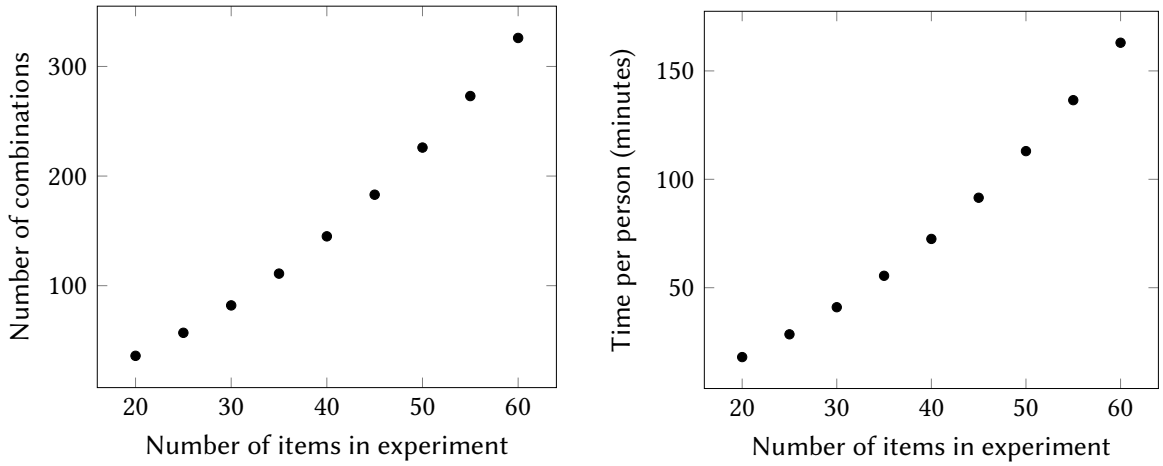
Table 9: Average number of seconds per task and group

	Group 1 (interjections)		Group 2 (verbs)		Group 3 (adverbs)	
	CS	Direct	CS	Direct	CS	Direct
Expert 1	217	≈ 90	225	≈ 90	148	≈ 90
Expert 2	155	≈ 15	129	≈ 15	117	≈ 15
Expert 3	156	≈ 20	199	≈ 20	327	≈ 20

Table 10: Observed (crowdsourcing, CS) and reported (direct) times for experts for the two modes of annotation, in minutes

crowdsourcing experiment depends on the number of items that make up the experiment. Thus, for 20 items and 4 items per task, if one calculates with a mean response time of 30 seconds per task, it would take three crowdsourcers approximately 18 minutes each if one were to collect three votes per task.[12] Figure 8a shows the number of combinations in the experiment when varying the number of items from 20 to 60 in increments of 5. Figure 8b shows the amount of time it would take each person on average to complete the project under the above constraints. It can be noted that there seems to be a curvilinear relationship between the number of items and the number of combinations; this relationship would be exponential if it were not for the redundancy-reducing algorithm used. If one looks at the time per person in relation to the number of items, the relation seems to mimic the relation between number of items and number of combinations. Further, doubling the number of crowdsourcers (from 3 to 6) leads to a reduction of time per crowdsourcer by half: for 20 items and 4 items per task with a mean response time of 30 seconds per task, it would take six crowdsourcers approximately 9 minutes each if one were to collect three votes per task.

[12] If one wants to collect three votes per task, the minimum required number of participants is three, as no (registered) participant will be shown the same task twice.



(a) Number of combinations

(b) Time per person

Figure 8: Number of combinations and time per person with varying number of items

7 Discussion

Among the burning questions in emerging crowdsourcing projects – within the domain of language learning – the three methodological questions below remain the most important at the current stage of development:

1. Who can be the crowd – with regards to the *background* of crowdsourcers?
2. How can *reliable* annotations be achieved with regards to design, number of answers and number of contributors? and
3. How should the *the results be interpreted* with regards to both research and practice?

The biggest gap that we have tried to fill with this study concerns the first (1) question, i.e. whether crowdsourcing as a method in language learning – within a limited domain of L2 resource annotation – could be used without explicit control for the background of the crowd.

Our results convincingly show that non-experts can perform on par with experts. We have seen that crowds with different backgrounds agree very well with each other, in comparison to previous research where CEFR raters of essays have often reached fairly low agreement. In fact, a mixed background crowd reaches "average" rankings faster. Note here that these conclusions are true of annotation carried out in a *comparative judgment* or *best-worst scaling* setting whereas previous work on essay rating has been done based on scales such as the CEFR-scale similar to our direct-labeling experiment with the CEFR-experts. To further confirm our findings, similar experiments need to be repeated for other languages, for other types of problems (e.g. annotation of texts for difficulty/readability), and for other subproblems of a given problem (e.g. annotation of single vocabulary items for difficulty). Similar conclusions have been made in projects within citizen science, among others by Kullenberg and Kasperowski (2016) where the experimental setup was not necessarily "comparative" in character. This leaves room for further experimentation.

In relation to question (2) the *reliability* of annotations, we have seen how the design of an annotation task influences the results. Clearly, a more traditional method

of annotation – using expert judgments – compares negatively to crowdsourced comparative judgments/best-worst scaling rankings. We have seen that experts do not agree with themselves when using comparative judgments versus categorical judgments, whereas the comparative judgment setting leads to homogeneous results between all groups of crowdsourcers regardless of their background, as shown in Section 2. According to Hovy and Lavid (2010) reliability of annotation of language resources has two types of major consequences, namely theoretical ones for shaping, extending and re-defining theories, and practical ones for use in the classrooms, but also in teaching and assessment practices. Unreliably annotated data can lead to biased – if not erroneous – theoretical conclusions and generalizations, as well as influence teaching and assessing practices in unwanted ways, as discussed among others in Carlsen (2012).

The above-stated *theory–practice* dichotomy can be traced down to the proficiency dimension of the MWE items in our experiment. On the one hand, the ranked list represents Multi-Word Expressions according to their difficulty from the learner's point of view and can thus be assumed to reflect stepwise development of their phraseological competence, which is of immediate interest to theoretical studies on L2 development. On the other hand, the scale represents perceptions of L2 professionals and – hypothetically – reflects their reasoning about what to teach/assess and in which order to do so based on their practical experience from teaching and assessing language learners, and has an immediate relevance for practical applications in "real life", including use in automatic solutions for language learning. It is very encouraging to see that the two perspectives (theoretical-developmental and practical) produce similar results and are so much in agreement with each other. However, this harmony can be observed only as long as we view vocabulary development as a continuum as opposed to groups of items belonging to one of a number of categorical proficiency levels.

In fact, both dimensions – theoretical and practical – are equally important. To understand how to teach and what to teach (practical dimension), we need to understand how learning is happening and (among others) observe which linguistic and cognitive aspects develop and in which order. While the produced scales give us material to study development of phraseology from a theoretical point of view, it is not obvious how to apply these scales to practical use (question (3) above) in teaching, assessment and Intelligent CALL, where categorical representations of proficiency are more customary and readily applicable. There are no indications in our crowdsourcing results as to where to draw the line between one level of proficiency and the other. We are not unique in facing these troubles, even though in other areas it can be a vice versa case:

> A weakness in this line of work is that SLA researchers have most often chosen to treat proficiency as a categorical variable and then have assessed mean differences in complexity values across proficiency groupings. Yet, this practice of converting interval variables (i.e. individual proficiency scores of some kind) into categorical ones (i.e. participants grouped by nominal proficiency levels) has always been criticized by statisticians because it discards much useful information. More specifically, it does away with the

> variance of continuous scores and leads to unreliability and increased likelihood of Type II errors (e.g. Troncoso Skidmore and Thompson, 2010), that is, the problem of failing to detect a difference, relationship, or effect that is in fact present because of some psychometric methodological problem, such as lack of power or (in the case at hand) lack of variance in the observations. It would be profitable in future work, therefore, to accumulate evidence from designs where both complexity and proficiency are treated as interval scales.
>
> (Ortega, 2012, p. 131)

This is a current challenge that needs to be addressed in the future (e.g. Paquot et al., 2020). Proficiency levels are always rather arbitrary (Hulstijn et al., 2010) as is also noted by the authors of CEFR (Council of Europe, 2001) who caution that "any attempt to establish 'levels' of proficiency is to some extent arbitrary, as it is in any area of knowledge or skill. However, for practical purposes it is useful to set up a scale of defined levels to segment the learning process for the purposes of curriculum design, qualifying examinations, etc." (p. 17). To summarize this part of the discussion, we view our results as a strong argument for treating vocabulary development as a continuum, while we also recognize the need to establish ways to partition vocabulary by levels of proficiency where these items can be taught.

On the practical side of crowdsourcing, our results show that a good and reliable agreement within a mixed crowd can be reached with two to three votes per task by at least three different voters. Considering these results, it might be interesting to use the same methodology for essay grading, especially since results from various experiments which have looked at inter-rater agreement in marking essays according to categorical proficiency levels have been less promising (cf Carlsen, 2012; Díez-Bedmar, 2012).

One of the limitations of the current setup lies in the use of the combinatorial algorithm which we apply to calculate the task pairings. As stated, we only achieve 77% *non-redundant* combinations, which means that certain pairs of expressions are included more than once and thus get more votes than other combinations, which might skew the picture. More involved statistical methods such as balanced incomplete block design (BIBD) (Yates, 1936) can be used to circumvent this problem. However, such methods impose hard constraints on the number of items and the number of items per task and not all combinations of number of items and items per task are able to satisfy these constraints. To the best of our knowledge, there exists no solution to the BIBD constraints for 60 items with 4 items per task.

The two methods of annotation – crowdsourcing by unknown crowd versus annotation by approved experts – have different dimensions of pros and cons. Here we have seen that time versus reliability can outweigh each other. In addition, one needs to consider that when using crowdsourcing, one has little control over the participant group and the time. Hence, neither method is superior on all accounts, but both are appropriate as long as one is aware of their weaknesses and strengths. If one is able to pay CEFR experts, one may get faster results. However, as seen in this study, one would need a large number of experts to reach consensus. Thus, expert knowledge can be fast and reliable *if* a large enough number of experts is consulted, to counteract the bias of individual

subjective opinions, but it is also expensive. If one does not have access to experts for various reasons, one can use crowdsourcing as an alternative to derive a relative ranking of expressions. The resulting ranking is similar regardless of whether one uses non-experts or experts, thus one may be able to realize such an experiment with non-experts only. In contrast to using experts for direct annotation, crowdsourcing is cheap however it takes longer time, both regarding the implementation and the actual crowdsourcing phase. Furthermore, with the set up we have chosen, one does not get concrete CEFR levels but rather a relative ranking. This data can, however, potentially be partitioned into more or less discrete proficiency levels by various techniques, should one desire to do so. The exploration and experimentation in this direction is future work.

8 Conclusion

In this study, we asked whether it influences the results in a crowdsourcing experiment aimed at ranking MWEs by difficulty if crowdsourcers are experts (L2 Swedish professionals) or non-experts (L2 Swedish learners / speakers). We set up different crowdsourcing experiments for the different target groups so as to be able to compare the results of different groups. The presented experiment suggests that it does not matter for this type of experiment if the crowdsourcer is an L2 speaker or an L2 professional, as the results produced by L2 speakers of Swedish, teachers of Swedish and CEFR experts are highly correlated. Concerning the design of the annotation task, we have convincingly shown that comparative design is a winner in contrast to explicit labeling: one does not need to have recourse to expert knowledge, and the results are much more homogeneous.

Furthermore, we explored how the number of votes influences the results and we found that with only two votes, the difference in results on a scale 1-60 is insignificant in comparison to three votes. Additionally, we found that sampling from a mixed-background group tends to produce more stable results. Indeed, using a mixed crowd produces results similar to results obtained from only expert annotations. This finding can further speed up crowdsourcing projects, since one can gather data with only one experiment instead of having to set up three distinct experiments for each target background. We also found that L2 *proficiency*, as measured by L2 professionals, does seem to correlate with L2 *development*, collected through intuitive judgments by L2 speakers.

These findings suggest that crowdsourcing might be a viable method to create a ranking of expressions by difficulty even in the absence of gold standard data. Our results suggest that there is a strong incentive in exploring crowdsourcing for other languages (if getting a scale is sufficient). For any new language and new item combination, we would suggest that the best-worst method be applied. There are reasons to believe that having strong "anchor words" for levels, i.e. words for which one knows the level with reasonable certainty, among the data can help create clusters around those with suggestions where to draw the line between one level and another, if there is a need for the pedagogical, assessment, CALL or other uses.

Future studies could investigate whether the same methodology produces the same results when applied to, for example, single word expressions or essays. Another direction for future research, as shortly mentioned above, might be how to partition an unordered, unlabeled set of expressions into different proficiency levels based, for example, on

clustering results. This might be achieved by adding certain *anchor* expressions to the experiment, i.e. expressions of which one knows with a sufficient degree of certainty their true label (i.e. target level). As a possible starting point, one could take the easiest and the hardest expressions overall from a ranking experiment such as the one presented, as the agreement at the extremes (very easy and very hard expressions) tends to be much higher than in the middle of the scale. Further, one might want to investigate how core and peripheral vocabulary can be identified based on different kinds of annotations.

References

Aker, Ahmet, Mahmoud El-Haj, M-Dyaa Albakour, Udo Kruschwitz, et al. 2012. Assessing Crowdsourcing Quality through Objective Tasks. In *LREC*, pages 1456–1461. Citeseer.

Alfter, David, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 1–7. Linköping University Electronic Press.

Alfter, David, Therese Lindström Tiedemann, and Elena Volodina. 2020. Expert judgments versus crowdsourcing in ordering multi-word expressions. *Swedish Language Technology Conference (SLTC)*.

Bachman, Lyle F, Adrian S Palmer, and Adrian S Palmer. 2010. *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press Oxford.

Bachman, Lyle F et al. 1990. *Fundamental considerations in language testing*. Oxford university press.

Benigno, Veronica and John de Jong. 2019. Linking vocabulary to the CEFR and the Global Scale of English: A psychometric model. *Developments in language education. A memorial volume in honour of Sauli Takala*, pages 8–29.

Bloomfield, Leonard. 1935. Linguistic aspects of science. *Philosophy of science*, 2(4):499–517.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18.

Borin, Lars, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Brezina, Vaclav and Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*, 36(1):1–22.

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling. 2016. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In Alan E Kazdin, editor, *Methodological issues and strategies in clinical research*. American Psychological Association.

Cambridge University Press. 2015. English Vocabulary Profile. https://www.englishprofile.org/wordlists. Accessed: 2019-11-11.

Capel, Annette. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1.

Capel, Annette. 2012. Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, 3.

Carlsen, Cecilie. 2012. Proficiency level—a fuzzy variable in computer learner corpora. *Applied Linguistics*, 33(2):161–183.

Cavnar, William B, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, volume 161175. Citeseer.

Chamberlain, Jon, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade, and Massimo Poesio. 2013. Using games to create language resources: Successes and limitations of the approach. In *The People's Web Meets NLP*, pages 3–44. Springer.

Chamberlain, Jon, Massimo Poesio, and Udo Kruschwitz. 2008. Phrase detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pages 42–49.

Chomsky, Noam and Morris Halle. 1965. Some controversial questions in phonological theory. *Journal of linguistics*, 1(2):97–138.

Chrzan, Keith and Megan Peitz. 2019. Best-Worst Scaling with many items. *Journal of choice modelling*, 30:61–72.

Čibej, Jaka, David Alfter, Iztok Kosem, and Elena Volodina. In preparation. Multiword expressions and language learning: validity of a crowdsourcing approach.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. www.coe.int/lang-cefr. Accessed 18.06.2020.

Culbertson, Gabriel, Solace Shen, Erik Andersen, and Malte Jung. 2017. Have your cake and eat it too: Foreign language learning with a crowdsourced video captioning system. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 286–296.

Díez-Bedmar, María Belén. 2012. The use of the common european framework of reference for languages to evaluate compositions in the english exam section of the university admission examination. *Revista de Educación*, 357:55–79.

Dürlich, Luise and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Eskildsen, Søren W. 2009. Constructing another language—usage-based linguistics in second language acquisition. *Applied linguistics*, 30(3):335–357.

EU Commission. 2016. General data protection regulation. *Official Journal of the European Union*, 59:1–88.

Forsberg Lundell, Fanny. 2020. Krävande krav. Vad ska språkkrav vara bra för?

Fort, Karën. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Fort, Karën, Bruno Guillaume, and Hadrien Chastant. 2014. Creating Zombilingo, a Game With A Purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, pages 2–6.

François, Thomas and Barbara De Cock. 2018. ELELex: a CEFR-graded lexical resource for Spanish as a foreign language. In *PLIN Linguistic Day 2018: Technological innovation in language learning and teaching*.

François, Thomas, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *LREC*, pages 3766–3773.

François, Thomas, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 213–219.

Garcia, Ignacio. 2013. Learning a language for free while translating the web. does duolingo work? *International Journal of English Linguistics*, 3(1):19.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. International corpus of learner English.

Habibi, Hanieh. 2019. LARA Portal: a Tool for Teachers to Develop Interactive Text Content, an Environment for Students to improve Reading Skill. In *Proceedings of the 12th annual International Conference of Education, Research and Innovation*, pages 8221–8229.

Hawkins, John A and Luna Filipović. 2012. *Criterial features in L2 English: Specifying the reference levels of the Common European Framework*, volume 1. Cambridge University Press.

Hovy, Eduard and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation*, 22(1):13–36.

Hulstijn, Jan H, J Charles Alderson, and Rob Schoonen. 2010. Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, pages 11–20.

Jiang, Yuchao, Daniel Schlagwein, and Boualem Benatallah. 2018. A Review on Crowdsourcing for Education: State of the Art of Literature and Practice. In *PACIS*, page 180.

Kilgarriff, Adam, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

Kuhn, Tanara Zingano, Peter Dekker, Branislava Šandrih, Rina Zviel-Girshin, Špela Arhar, and Tanneke Schoonheim Holdt. 2019. Crowdsourcing corpus cleaning for language learning resource

development. In *EUROCALL Conference 2019*, page 163.

Kullenberg, Christopher and Dick Kasperowski. 2016. What is citizen science?–a scientometric meta-analysis. *PloS one*, 11(1):e0147152.

Lafourcade, Mathieu and Alain Joubert. 2008. JeuxDeMots: un prototype ludique pour l'émergence de relations entre termes. In *JADT'08: Journées internationales d'Analyse statistiques des Données Textuelles*, pages 657–666.

Lafourcade, Mathieu, Alain Joubert, and Nathalie Le Brun. 2015. *Games with a Purpose (GWAPS)*. Wiley Online Library.

Leńko-Szymańska, Agnieszka. 2015. The english vocabulary profile as a benchmark for assigning levels to learner corpus data. *Learner corpora in language testing and assessment*, pages 115–140.

Lesterhuis, Marije, San Verhavert, Liesje Coertjens, Vincent Donche, and Sven De Maeyer. 2017. Comparative judgement as a promising alternative to score competences. In *Innovative practices for higher education assessment and measurement*, pages 119–138. IGI Global.

Lindström Tiedemann, Therese, Daniela Piipponen, Beatrice Silén, David Alfter, and Elena Volodina. In preparation. Multi-word Expressions in Swedish as a second language – typology and initital results.

Louviere, Jordan J, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Lyding, Verena, Lionel Nicolas, Branislav Bédi, and Karën Fort. 2018. Introducing the European NETwork for COmbining Language LEarning and Crowdsourcing techniques (enetcollect). *Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL*, 2018:176–181.

Nicolas, Lionel, Verena Lyding, Claudia Borg, Corina Forascu, Karën Fort, Katerina Zdravkova, Iztok Kosem, Jaka Čibej, Špela Arhar Holdt, Alice Millour, Alexander König, Christos Rodosthenous, Federico Sangati, Umair ul Hassan, Anisia Katinskaia, Anabela Barreiro, Lavinia Aparaschivei, and Yaakov HaCohen-Kerner. 2020. Creating expert knowledge by relying on language learners: a generic approach for mass-producing language resources by combining implicit crowdsourcing and language learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 268–278, Marseille, France. European Language Resources Association.

O'Muircheartaigh, Colm, George Gaskell, and Daniel B Wright. 1995. Weighing anchors: Verbal and numeric labels for response scales. *Journal of official statistics*, 11:295–308.

Ortega, Lourdes. 2012. Interlanguage complexity. *Linguistic complexity: Second language acquisition, indigenization, contact.*, Vol.13.

Paquot, Magali, Hubert Naets, and Stefan Th Gries. 2020. Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb + object structures in LONGDALE. *Learner Corpus Research Meets Second Language Acquisition*, page 122.

Pilán, Ildikó, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a

linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111.

Snow, Rion, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Solemon, Badariah, Izyana Ariffin, Marina Md Din, Rina Md Anwar, et al. 2013. A review of the uses of crowdsourcing in higher education. *International Journal of Asian Social Science*, 3(9):2066–2073.

Spinner, Patti and Susan M Gass. 2019. *Using judgments in second language acquisition research*. Routledge.

Stegbauer, Christian, Elisabeth Bauer, Elisabeth Kartashova, and Alexander Rausch. 2009. *Wikipedia*. Springer.

Stemle, Egon W., Adriane Boyd, Maarten Jansen, Therese Lindström Tiedemann, Nives Mikelić Preradović, Alexandr Rosen, Dan Rosén, and Elena Volodina. 2019. Working together towards an ideal infrastructure for language learner corpora. In Andrea Abel, Aivars Glaznieks, Verena Lyding, and Lionel Nicolas, editors, *Widening the Scope of Learner Corpus Research*, Corpora and Language in Use, pages 427–468. Presses universitaires de Louvain, France.

Sweedler-Brown, Carol O. 1985. The influence of training and experience on holistic essay evaluations. *The English Journal*, 74(5):49–55.

Tack, Anaïs, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.

Tenfjord, Kari, Paul Meurer, and Knut Hofland. 2006. The ASK Corpus-a Language Learner Corpus of Norwegian as a Second Language. In *LREC*, volume 6, pages 1821–1824.

Troncoso Skidmore, Susan and Bruce Thompson. 2010. Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement*, 70(5):777–795.

Volodina, Elena, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 128–144.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 206–212.

Volodina, Elena, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016b. SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 76–84. Linköping University Electronic Press.

Weigle, Sara Cushing. 1998. Using facets to model rater training effects. *Language testing*, 15(2):263–287.

West, Michael Philip. 1953. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans, Green.

Yates, Frank. 1936. Incomplete randomized blocks. *Annals of Eugenics*, 7(2):121–140.